International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

The Possibility of Selective Skin Lesion Classification in Convolutional Neural Networks

Evelyn Anyebe*

Postgraduate student University of Dundee,Dundee, DD1 4HN, United Kingdom

Email: eanyebe@dundee.ac.uk

Abstract

Selective classification of skin lesion images and uncertainty estimation is examined to increase the adoption of convolutional neural networks(CNNs) in automated skin cancer diagnostic systems. Research on the application of deep learning models to skin cancer diagnosis has shown success as models outperform medical experts [1]. However, concerns on uncertainty in classifiers and difficulty in approximating uncertainty has caused limited adoption of CNNs in Computer-aided diagnostic systems (CADs) in health care. This research propose selective classification to increase confidence in CNN models for skin cancer diagnosis. The methodology is based on SoftMax response(SR), MC dropout and risk-coverage performance evaluation metric. Risk-coverage curves gives physicians and dermatologist information about the expected rate of misclassification by a model. This enable them to measure the reliability of the classifier's predictions and inform their decision during skin cancer diagnosis. MC dropout uncertainty estimate was shown to increase accuracy for Melanoma detection by 1.48%. The proposed selective classifier achieved increase melanoma detection. The sensitivity of melanoma increased by 9.91% and 9.73% after selective classification at a coverage of 0.7. This study showed that selective classification and uncertainty estimation can be combined to promote adoption of CNNs in CADs for skin lesions classification.

Keywords: skin lesion classification; selective classification; deep learning; uncertainty estimation; convolutional neural network(CNNs); MC dropout.

---

* Corresponding author.

1. Introduction

A CNN classifier trained on skin lesion images opens many avenues for automatic skin cancer diagnosis - self-diagnosis on mobile devices, CADs for use by clinicians and dermatologists among others. CNNs are deep neural networks that have remarkable ability to detect patterns and extract features in images; thereby leading to extensive research on them for skin lesion classification [1,2,3,4]. These research works have reported CNNs performing better than physicians and dermatologist at skin lesion classification [1,5] but, - these models are yet to be extensively utilized in CADs for hospitals or apps for patients. This is due to distrust in the reliability of deep learning models and invariably CNNs due to their black box nature [2]. Explaining the ability of CNNs to generalize well and correctly classify images is difficult. Machine learning experts find it hard to determine when they fail. CNNs produce overconfident outputs which makes it difficult to measure uncertainty in their predictions. Even when given images that are unrelated to the image domain on which the CNNs are trained, the models still make predictions with high confidence. A classifier for skin lesions should not make predictions on images that are not related to skin lesions. For instance, image of a person's hair, healing cut or scar, skin partly covered by clothing among others. The CNN should tell when it doesn't know, reject the prediction and guide the physician on the predictive uncertainty through its output. This is selective classification or reject option[2,6,7]. On the other hand, training a machine learning model approximates the true distribution of the data generation process. It is dependent on amount of data available and leads to existence of predictive uncertainty [8]. The de-facto and accepted method for uncertainty estimation in machine learning is Bayesian inference[9,10] but in practice other methods such as MC dropout [2,10,11] and statistics of ensemble are often used for deep neural networks. Bayesian inference is intractable in CNNs due to large number of parameters being learnt. As such the overconfident predictions and difficulty of uncertainty estimation in CNNs has limited their adoption for skin cancer diagnosis. This study examines the possibility of selective classification of skin lesions in order to increase the use of CNNs in for skin cancer diagnosis.

1.1 Sources of skin lesion images

Public datasets of skin lesion images are limited hence the need to discuss their sources. Dermoscopic and clinical images are two types of skin lesions images commonly used in machine learning. Clinical images are cheap to obtain and important for creating large datasets while dermoscopy is the standardized method for obtaining skin lesion images. Dermoscopic images may also include images obtained by other non-invasive optical technologies like optical coherence tomography. Dermoscopy imaging magnifies the morphology of skin lesions and models trained on such images to achieve better accuracy. On the other hand, models trained on clinical images show lower performance when evaluated. With clinical images, performance is better in binary classification (benign vs malignant) compared to multi-class classification [3]. Using very good cameras to obtain clinical images often help accentuate features. With advances in technology, mobile devices with high capacity cameras can obtain these images. In turn, models can be built for these devices and utilized for self-diagnosis. The International Skin Imaging Collaboration (ISIC) has made public datasets such as HAM10000 available for research. ISIC archive contains the largest publicly available dataset of dermoscopic images [12]. ISIC 2019 dataset is utilized in this study. It was provided for ISIC challenge 2019 with title "towards melanoma detection". As such this study also reports on the model's ability to detect Melanoma skin lesions.

Worth noting is class imbalance in the dataset. This is because incidence of melanoma is very rare[13,14]. It constitutes 10% of skin lesion while non-melanoma skin lesions constitute 90% [15]. Stratified random sampling is used to split the dataset.

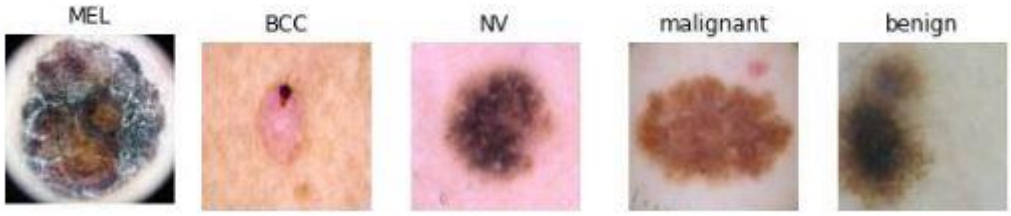


Figure 1: Examples of Skin lesion images in ISIC 2019 dataset

ISIC 2019 dataset consist of 25,331 skin lesion images and metadata files. The images are categorized into 'MEL', 'NV', 'BCC', 'AK', 'BKL','DF', 'VASC' and 'UNK' corresponding to '0', '1', '2', '3','4', '5', '6' and '7'. In the ISIC challenge of 2019, Ensemble of multi-RES EfficientNets, ensemble of EfficientNet B3-B4 and DenseNet 161 were the top two networks and they attained 63.3% and 60.7% balanced multi-class accuracy (according to ISIC leader board).

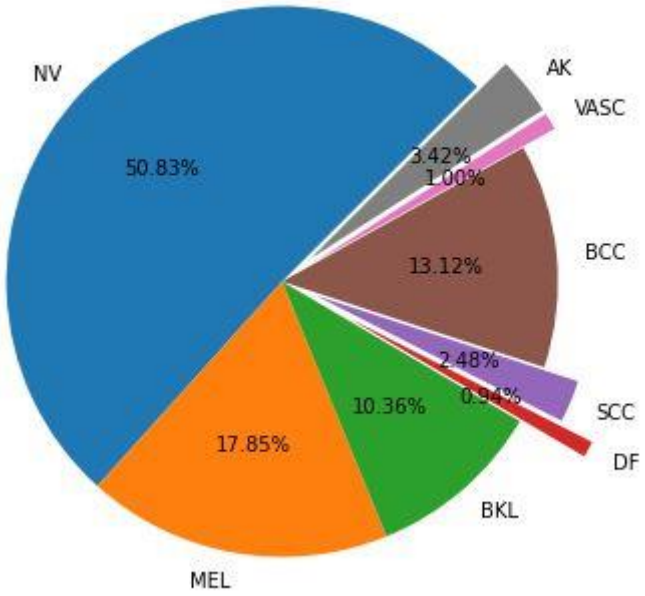


Figure 2: Pie chart of ISIC 2019 dataset and the proportion of the classes

1.2 Related works in uncertainty estimation

Bayesian Neural Networks(BNNs) has been explored for uncertainty in deep learning with many algorithms[8,10,16]. BNNs were proposed by Denker and LeCun 1991, Mackay 1992, Neal, 2012 [9,10,17] to

capture uncertainty in deep neural networks. In these works, each weight is represented by a probability distribution (Gaussian) instead of a real number and learning is done by computing the posterior $p(w|D)$. Calculating the posterior on the weights could not be done analytically.-. therefore, BNNs needed approximation methods. Two broad approximation methods, Markov Chain Monte Carlo (MCMC) and Variational inference [10] are used. A MCMC sampling method using Hamilton Dynamics was introduced by Neal in 1993 [9]. This method produces a sample of parameters without directly calculating the posterior. This proved impractical in CNNs because MCMC runs on the whole dataset [10]. Also, sampled parameters from the posterior after each iterative step needed to be stored. Other MCMC sampling methods were used such as Stochastic MCMC but they still required storing of sampled parameters which is very inefficient for CNNs [9]. Variational inference is a standard in BNNs. The first application of Variational inference was in the paper by Grave, 2011 where MC sampling was used to estimate the posterior distribution [9,10,17]. However, this method performed badly due to lack of correlation between the weights [10]. In 2015, Blundel et.al provided an improvement by placing Gaussian distribution for the posterior distribution [10]. Variational inference approximates the posterior distribution by a tractable variational distribution $q_\theta(w)$ indexed by variational parameter $\theta \epsilon \Theta$ for variational parameter space Θ. As such, the optimal variational distribution is the closest distribution to the posterior among the pre-determined family $Q = \{q_\theta(w) : \theta \epsilon \Theta\}$. Blundel et. al parameterizing $q_\theta(w)$ as a Gaussian doubles the number of parameters due to mean and variance which makes inference more challenging. Reference [10] proposed MC dropout - where gradient-based optimization procedure on a dropout neural network was shown as equivalent to a variational approximation $q_\theta(w)$ on a BNN. Another approach that is used but not based on Variational inference is ensemble of models [10]. Here, each model produces a point estimate rather than a distribution. Many randomly initialized instances of a model are trained on the same dataset or subsets of the dataset. The variance of outputs from all models is then estimated for a given test input [9,10,11] (other statistical measures may be employed as well). Ensemble of models is based on statistical modelling and variation in the outputs from many models reflects uncertainty. Ensemble methods require more memory for storing weights of the various models. On the other hand, MC dropout require the network to perform hundreds of feed forward iteration for a single prediction, this can increase latency.

1.3 Related works in selective classification

Selective classification[1] also called reject option has been studied in machine learning algorithms such as SVM, decision trees and nearest neighbours [6]. In CNNs, selective classification is based on a confidence score [19] . Often, maximum SoftMax outputs of the network serve as confidence scores and a threshold is then carefully selected [19]. Confidence scores are also obtained using MC dropout [19]. Another procedure for reject function is based on ensemble of models such that the mean predicted values or predictive variance are used as confidence scores[6,19]. Jointly learning the reject function called integrated reject was first proposed by Cortes and his colleagues [7,20]. This idea was extended to Selectivenet in 2018 [6]. Selectivenet used risk-coverage metric to evaluate the selective model. In medicine, selective classification has been used for liver disease diagnosis [21], image segmentation [9] among others. For skin lesion classification, focus has been on uncertainty estimation [2,11].

[1] Selective classification is also referred to as reject option in this section

2. Materials and methods

This study adopts MC dropout for uncertainty estimation and carried out selective classification using SoftMax response with risk-coverage method. The neural network is based on EfficientnetB0 and ISIC 2019 dataset is employed for training and testing.

2.1 Definition of the selective model

Let X be images of skin lesions and Y be labels. Then $p(X,Y)$ is the distribution over $X \times Y$. A classifier f is defined as f(x) : X → Y has true risk $R(f)$ with respect to P.

$$R(f) \triangleq E_{p(x,y)}[l(f(x),y)] \quad (1)$$

Where $l: Y \times Y \rightarrow R^{+}$ is a given loss function.

Given an independent and identically distributed(i.i.d) random dataset $S_m = \{(x_i, y_i)\}_{i=1}^{m} \subseteq (X \times Y)^m$ sampled from *P(X ,Y)*, the classifier *f*(x) has an empirical risk $\hat{r}(f|S_{m)}$ defined by (2) below:

$$\hat{r}(f|S_{m)} \triangleq \frac{1}{m}\sum_{i=1}^{m} l(f(x_i),y_i) \quad (2)$$

A selective classification model is then defined as a pair (f, g)

$$(f,g)(x) \triangleq \begin{Bmatrix} f(x), & if\ g(x)=1; \\ don't\ know, & if g(x)=0. \end{Bmatrix}$$

Where $f(x)$ is a classification function, $g(x)$ is a reject function defined as $g(x): X \rightarrow \{0,1\}$ [6]. The selective classifier does not make prediction when g(x)=0. Thus, selection is a binary problem; accept (1) or reject (0) i.e. $g(x): X \rightarrow [0,1]$. Selection can be done probabilistically or with a threshold. The latter is used in this study. A threshold is obtained using coverage -where coverage ($\phi(g)$) is the probability mass of the non-rejected region in X.

$$\phi(g) \triangleq E_p[g(x)] \quad (3)$$

The performance of a Selective model can be measured by coverage and selective risk. Selective risk represents rate of misclassification after reject defined by equation (4)

$$R(f,g) \triangleq \frac{E_{p(x,y)}[l(f(x),y)]}{\phi(g)} \quad (4)$$

Therefore, the corresponding empirical coverage and risk for S_m are:

$$\phi(g|S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i) \quad (5)$$

$$\hat{r}(f|S_m) \triangleq \frac{\frac{1}{m}\sum_{i=1}^{m} l(f(x_i), y_i)\, g(x_i)}{\phi(g|S_m)} \quad (6)$$

The objective of the selective model becomes optimizing selective risk given a constraint on the coverage and vice-versa [6].

$$\theta^* = argmin_{\theta\epsilon\Theta}\big(R(f_\theta, g_\theta)\big) \quad (7)$$

$$s.t\ \ c \geq \phi(g_\theta)$$

for a specified coverage rate $0 < c \leq 1$, and hypothesis class θ(the set of parameters for a given CNN). Taking maximum SoftMax output to represents g(x) – the outputs are considered as confidence scores. These values with respect to a threshold are transformed to 0 or 1 (i.e. $g(x): X \rightarrow [0,1]$) using a reject function based on a coverage c. The threshold is obtained by ranking the distribution $g(x_{i)})$ based on the target coverage. Given a coverage c, threshold is set to $100(1-c)$ percentile of the distribution $g(x_{i)}$. For a given coverage, the selective risk involved is computed and shown using risk-coverage curves. Selective risk provides information needed for making a choice of coverage for deployment of a model. If the model is deployed, experts' (physicians and dermatologists) reliance on predictions will be guided by this information. Apart from risk-coverage curves, the selective model was evaluated using accuracy, confusion matrix and sensitivity in results section.

2.2 MC dropout

Dropout layers are used in many deep models for regularization by randomly dropping (i.e. setting activations to zero) units and are required for MC Dropout. A parameter drop rate: - *pdrop* is required by dropout procedure. The drop rate is the percentage of units dropped at the layer. By dropping units randomly in the CNN, dropout enable sampling of different CNN architecture. Given an unknown input, the predictive posterior is estimated by performing T stochastic iterations of feed forward pass in a neural network. MC Dropout is defined by equation (8).

$$p(y*|x*, X, Y) \approx \int p(y*|x*, w) q_\theta^*(w) dw \approx \frac{1}{T}\sum_{t=1}^{T} p(y*|x*, \widehat{w_t}) \quad (8)$$

Where $\widehat{w_t}$~q(w), $\widehat{w_t}$ are network parameters drawn from the network T number of times. With MC dropout, the network is required to perform hundreds of feed forward iteration for a single prediction. MC dropout results in an approximate predictive posterior distribution. The mean of the MC iterations represents the prediction of the network (9). Uncertainty is represented by the predictive variance as defined in equation (10).

$$\mu_{pred} \approx \frac{1}{T}\sum_{t=1}^{T} p(y*|x*, \widehat{w_t}) \quad (9)$$

$$H[y*|x*, D_{train}] := -\sum_{c} p(y*= c|x*, D_{train}) \log p(y*= c|x*, D_{train}) \quad (10)$$

The class with the largest predictive mean is selected as prediction and variance of the predictive distributions for each class represents uncertainty.

2.3 EfficientNetB0

EfficientNetB0 is utilized as the feature extractor for the CNN architecture and a classification head is added. EfficientNetB0 is the baseline network from which EfficientNets B1-B7 family of CNNs were derived. It was reported to have higher performance, decreased latency, top-1 and top-5 accuracy of 77.3% and 93.5% on ImageNet data compared to the high performing ResNet50 [22]. Therefore, it was a suitable choice for the CNN in this study.

2.4 Experiment

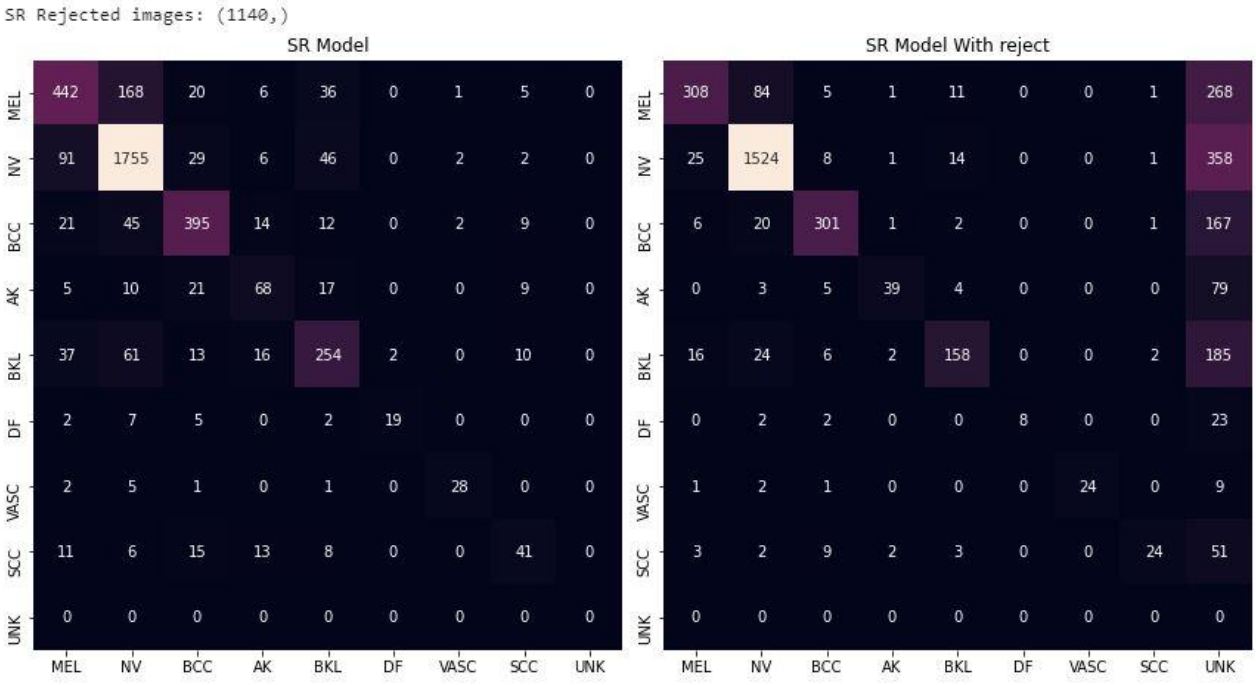
The model is implemented using TensorFlow 2.2 and Keras API. After model training, the experiment compared MC dropout and a single prediction representing SoftMax response. This is to examine the effect of variance as models' uncertainty. The experiments also examined MC variance and MC predictive distributions. EfficientNetB0 was obtained from https://github.com/qubvel/efficientnet and utilized as base model. The classification head contains flatten layer followed by a dropout, dense and batch normalization layers. The final dense layer has 9 neurons representing number of classes in ISIC 2019 dataset. The model used three dropout layers in the classification head. The first two dropout layers utilized 0.3 drop rate while the final dropout layer utilized 0.2 as drop rate. The model was optimized using SGD with learning rate=0.01, momentum=0.9 and nesterov=True and categorical cross entropy loss. A learning rate schedular call back reduced learning rate by 0.25 after 20 epochs. An early stopping call back monitored validation loss with min-delta = 0.0001 and patience = 3. As such the model was trained for 120 epochs. The images were split into train, test and validation sets with 17739, 3796, 3796 images respectively. The pixel values were normalized to range from 0 to 1 and the images were reshaped to 224x224x3. At inference time, T =100 iterations were made for each test sample. The mean of distribution is used for MC Dropout prediction. The variance was taken to represent uncertainty.

3. Results

The model was tested on 3796 test images and results are shown for SoftMax Response, MC Dropout among others.

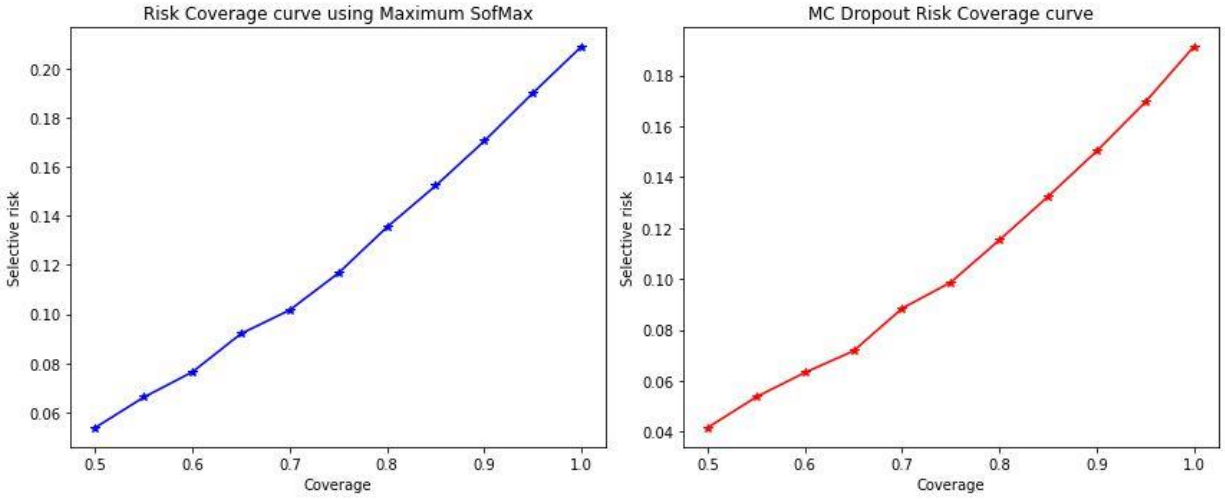
3.1 SoftMax response

After evaluation, the model achieved an accuracy of 79%$\pm$0.3 without rejection. The confusion matrix with and without reject is provided in figure (3). Also, the risk-coverage curve of the model is shown.



a) b)

Figure 3: Confusion matrix of model. a) shows confusion matrix of without rejection b)shows confusion matrix with rejection where coverage = 0.7

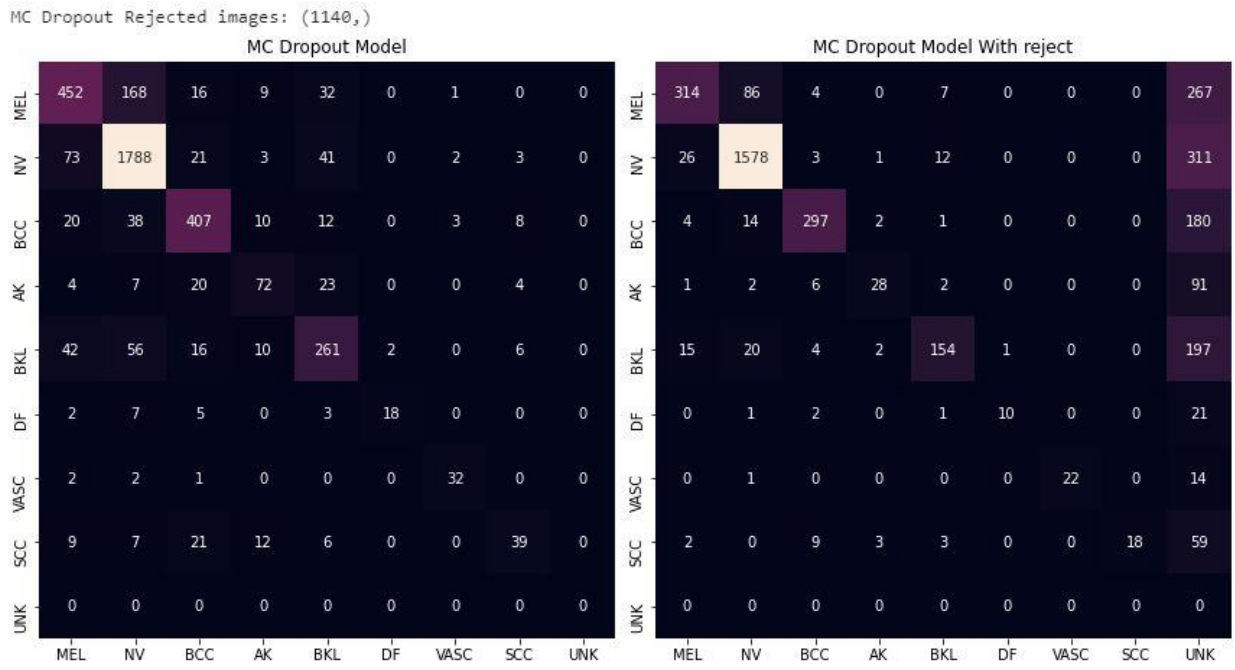


a) b)

Figure 4: Risk-Coverage curves a) shows risk-coverage for maximum SoftMax output as confidence scores b)shows risk-coverage using maximum MC dropout mean as confidence scores

3.2 MC Dropout

After performing MC Dropout using the test dataset, the model achieved 80%±0.2 accuracy. The risk-coverage curve and confusion matrices are shown in figure (4) and (5). Also, table (2) and (3) compares the risk coverage and melanoma sensitivity of SoftMax Response and MC Dropout. The predictive variance of MC Dropout is examined, and MC predictions are observed using histogram and scatter plot. Furthermore, predictive variance is used as confidence scores for selective classification. Predictive variance selective risk is compared to SR and MC dropout in table (1) and figure (6).



a) b)

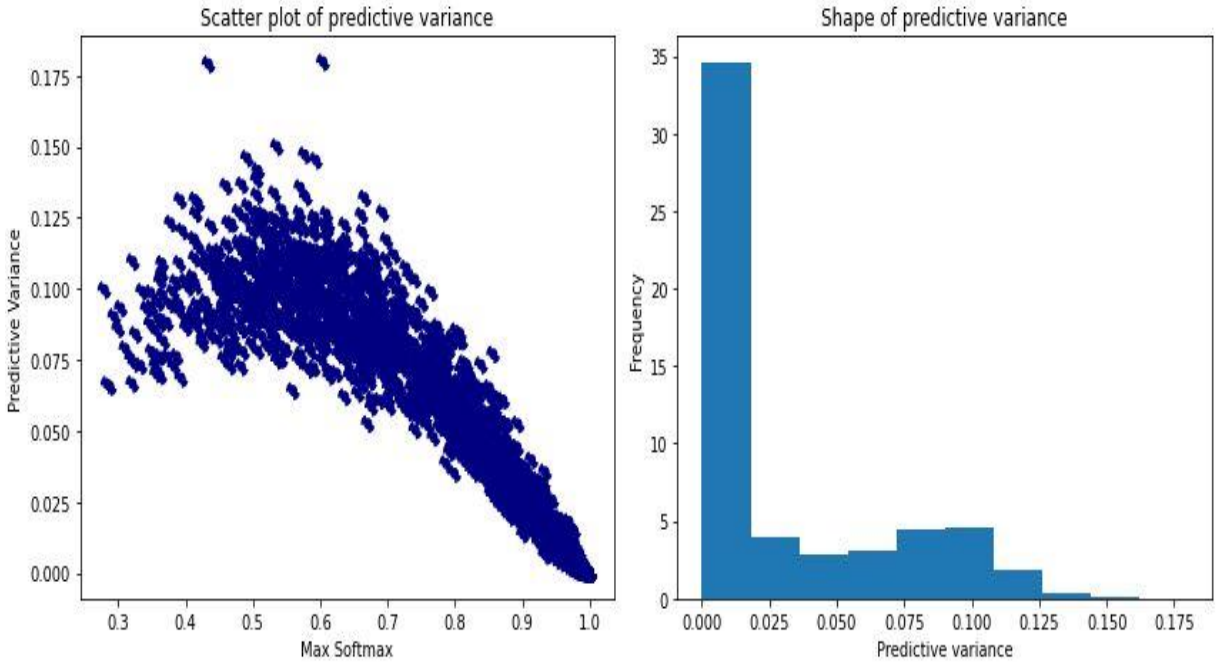
Figure 5: MC Dropout confusion matrices a) shows confusion matrix without rejection b) shows confusion matrix with rejection where coverage = 0.7

Table 1: Selective risk of SR and MC Dropout

Calibrated coverage	**SR selective risk**	**MC selective risk**	**Predictive variance**
0.7	0.10	0.09	0.09
0.75	0.12	0.10	0.10
0.80	0.14	0.12	0.12
0.85	0.15	0.14	0.14
0.90	0.17	0.15	0.15
0.95	0.19	0.18	0.18
1.0	0.20	0.19	0.19

Table 2: Comparison of Sensitivity(%) for Melanoma for SR and MC dropout

Melanoma Detection	SR	MC Dropout	MC Dropout -SR
Sensitivity without rejection	65.19	66.67	1.48
Sensitivity with rejection c=0.7	75.10	76.40	1.3
Sensitivity increase	9.91	9.73	



a) b)

Figure 5: Predictive variance a) shows a scatter of predictive variance and maximum SoftMax b) Histogram of predictive variance

Table 3: Statistical summary of predictive variance

Metric	Predictive variance value
Minimum	4.26e-16
Maximum	0.18
Mean	2.82e-02
Median	2.98e-03
25%	7.07e-06
75%	5.65e-02
Standard deviation	3.87e-02

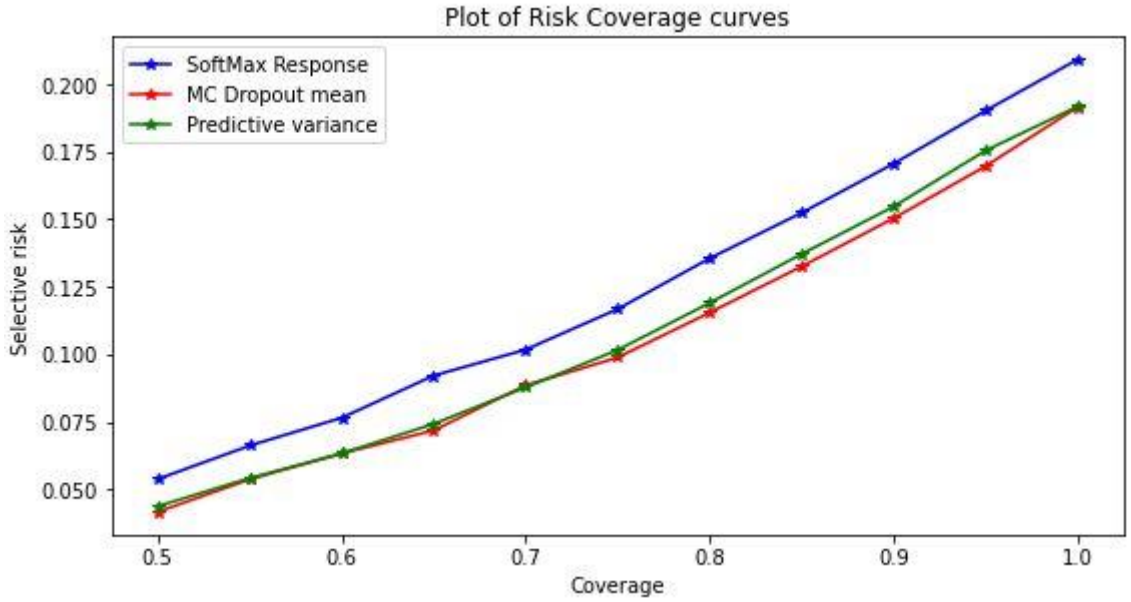
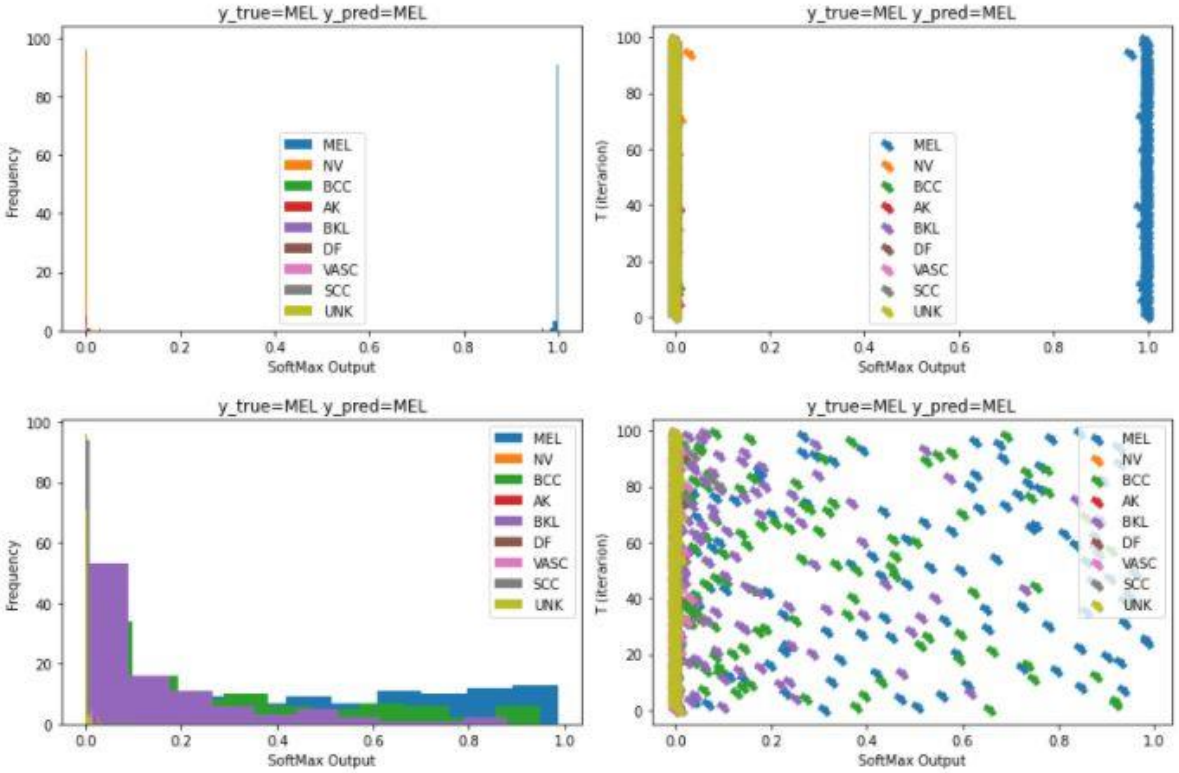


Figure 6: Risk coverage curves of SoftMax response, MC dropout mean, and predictive variance compared

75% of the predictive variance is approximately ≤ 0.05. This information though showing the predictive uncertainty is a single value and vague. Nonetheless, a clearer view is provided by the histograms and scatter plots of MC predictions.



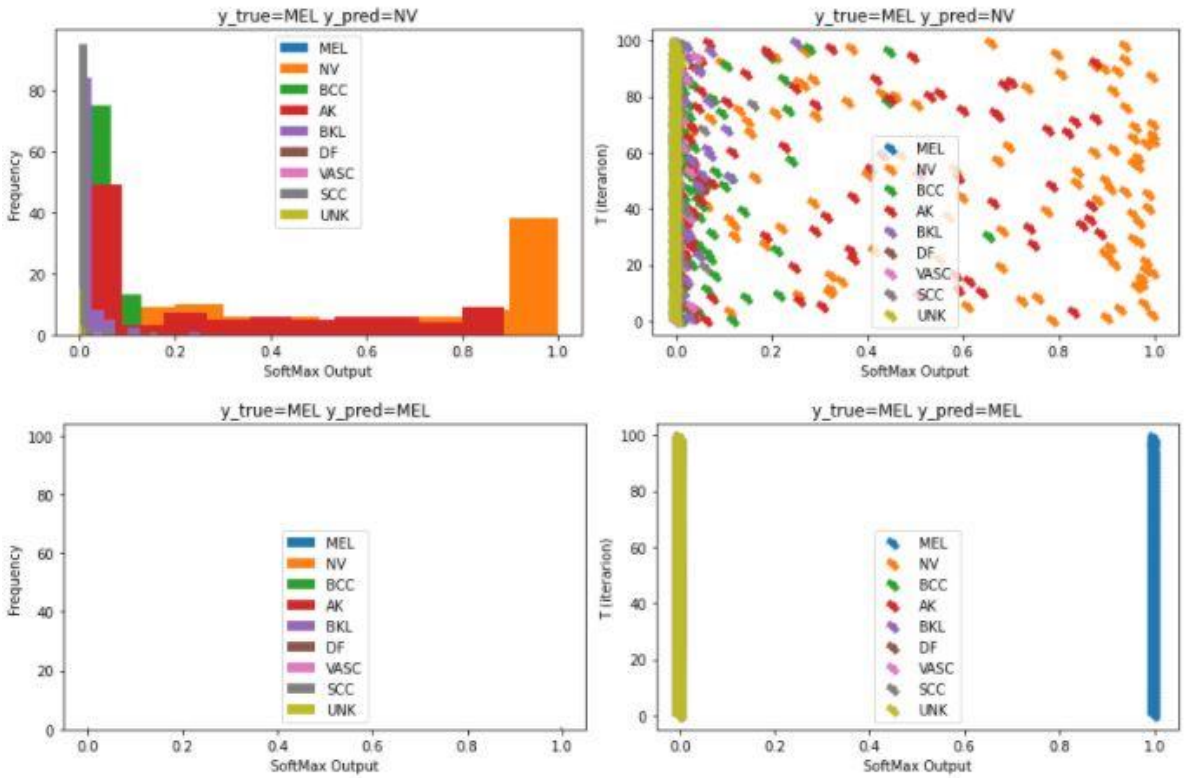
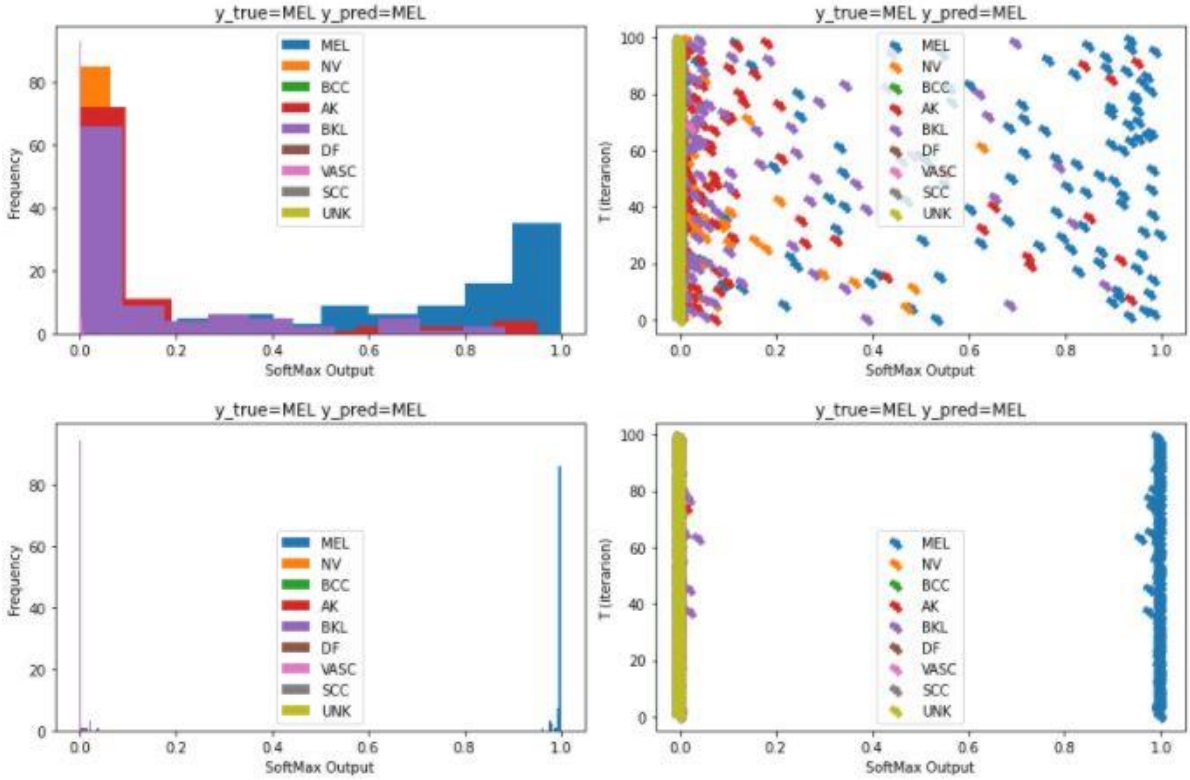


Figure 7: Histogram and scatter plot of MC predictions for four test examples



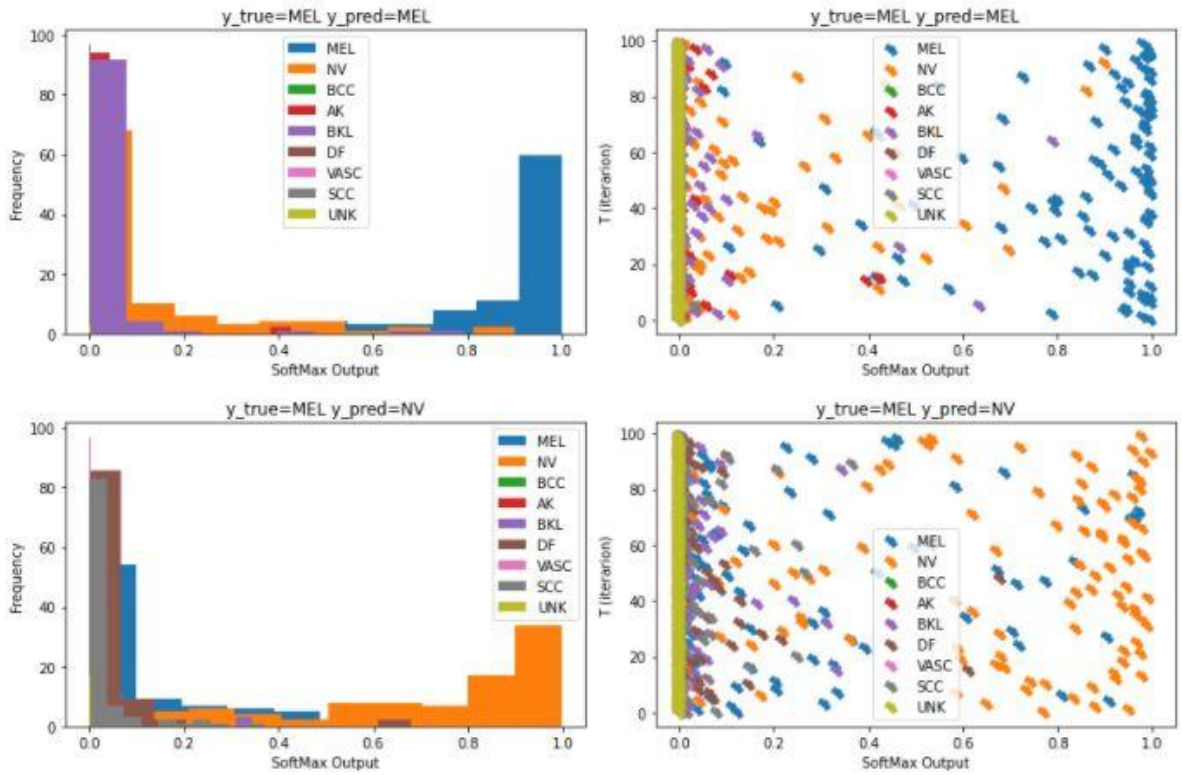


Figure 8: Histogram and scatter plot of MC predictions for four test examples

4. Discussion

From figure (3), it was observed that both true Melanoma and False negative melanoma images were rejected. However, Melanoma sensitivity increased by 9.91%. Similarly, in MC Dropout where argmax mean MC prediction is used as confidence scores, Melanoma sensitivity increased by 9.73%. The risk-coverage curves for SoftMax Response and MC Dropout showed a selective risk of approximately 0.2 (i.e. 20%) when all test predictions are accepted (coverage c=1.0). Depending on a tolerable risk, for instance at c=0.7, where there is a 10% selective risk, the threshold can be accepted as the setting for use in CADs. Observing the predictive variance and using it as confidence scores showed no significant reduction in selective risk over argmax MC mean. However, the statistics of predictive variance can guide physicians' decisions when accepting the model's prediction. Figure (7) and (8) showed histograms and scatter plots of MC predictions for eight test images. It was observed that only three graphs show a clear distinction between the predicted class and the other classes. As such, the graph of MC predictions can be used to guide physicians' confidence in the automated skin cancer diagnostic system after selective classification.

5. Conclusion

This study showed how selective classification of skin lesion images can be carried out using maximum SoftMax output or MC Dropout. It showed that risk-coverage curves can be used to select the threshold value. The distribution of MC Dropout predictions reflects the model's confidence in its prediction. Further work is recommended to reduce selective risk of the models such as integrated reject option. Use of out-of-distribution,

probabilistic or grid search method should be examined for threshold. A comparison of MC Dropout and ensemble method can be carried out.

Acknowledgement

With profound gratitude, I pay tribute to Petroleum Trust Development Fund (PTDF) for their immerse support. I am most grateful to my master thesis supervisor, CVIP group, lecturers and staff of computing, school of science and engineering, university of Dundee who contributed immensely to this work.

References

[1]. Esteva et al, "Dermatologist-level classification of skin cancer with deep nueral networks," Nature, vol. 542, pp. 115-217, 2017.

[2]. A. Mobiny, A. Singh, and H.V. Nguyen, "Risk-Aware Machine Learning Classifier for skin lesion diagnosis," Journal of clinical medicine, vol. 8, no. 8, p. 1241, 2019.

[3]. Y. Fujisawa, S. Inuoe, and Y. Nakamura, "The possibility of deep learning based computer-aided skin tumour classifier," Frontiers in medicine, vol. 6, pp. 1-10, 2019.

[4]. S. Chan et al, "Machine Learning in Dermatology: Current Applications, opportunities and limitations," Adis Journals, 2020.

[5]. N. Gessert et al, "Skin Lesion Classification Using Loss Balancing and ensemble of Multi-Resolution EfficientNets," in ISIC, 2019.

[6]. Y. Geifman, and R. El-Yaniv, "Selectivenet: a deep neural network with and integrated reject," 2019.

[7]. C. Cortes, G. DeSalvo, and M. Mohri, "Boosting with abstention," In Advances in Neural, pp. 1660-1668, 2016.

[8]. K. Murphy, in Machine learning: A probabilistic perspective, MIT Press, 2012.

[9]. Y. Kwon et al, "Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation".

[10]. Y. Gal, "Uncertainty in deep learning," University of Cambridge, Cambridge, 2016.

[11]. P.V Molle et al, "Quantifying Uncertainty of Deep Neural Networks in skin lesion classification," 2019.

[12]. ISIC, "Challenge2019.Isic-Archive," ISIC. (n.d.), 2019. [Online]. Available: https://challenge2019.isic-archive.com/. [Accessed 2020 05 2020].

[13]. Z. Apalla et al, "Skin cancer: Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approches," Demartol. Ther, pp. 7,5-19, 2017.

[14]. C. R. UK, "Melanoma skin cancer incidence statistics," 27 06 2020. [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics.

[15]. N. C. Institute, "Surveilance, Epidemology and end results program: Cancer stats," 25 06 2020. [Online]. Available: https://seer.cancer.gov/statfacts.

[16]. I. Goodfellow et al, Deep Learning, MIT Press, 2016.

[17]. E. Hullermeier, and W. Waegeman,, "Aleatoric and Epistemic Uncertainty in Machine," 2020. [Online]. Available: https://arxiv.org/abs/1910.09457. [Accessed 10 07 2020].

[18]. Y. Gal and Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," 2016.

[19]. Y. Geifman, and R. El-Yaniv, "Selective classification for deep neural networks," In Advances in neural information processing systems, no. b, pp. 4878-4887, 2017.

[20]. C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," In International Conference on Algorithmic Learning Theory, pp. 62-82, 2016.

[21]. K. Hamid, A. Asif, W. Abbasi, D. Sabih, and F. U. A. A. Minhas, "Machine Learning with Abstention for Automated Liver Disease Diagnosis," International Conference on Frontiers of Information Technology (FIT), Vols. 356-361, 2017.

[22]. M. Tan, and V. Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in ICML, California, 2019.